A SIMULATION STUDY OF VARIOUS ESTIMATORS WHICH USE AUXILIARY DATA IN AN ESTABLISHMENT SURVEY

Hyunshik Lee and James Croal, Statistics Canada
Hyunshik Lee, 11-J, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6

Key words: ratio estimator, regression estimator, Mickey's estimator, small area estimation, robust technique, correlation, linear model.

1. INTRODUCTION

The Survey of Employment, Payroll and Hours (SEPH) is a monthly establishment based survey conducted by Statistics Canada. It collects data on Total Employment, Earnings, Hours of paid employment, and other related variables. The objective of the survey is to measure both the monthly levels and the month-to-month changes of those variables, for all industries except agriculture, fishing and trapping, private household services, religious organizations and military services.

This study was prompted by two major corporate concerns:

- to reduce the respondent burden for small employers;
- to reduce the size of the monthly SEPH sample.

Employers are required to remit certain payroll deductions to Revenue Canada Taxation, at least once per month. These remittances consist of Canada Pension Plan premiums, Unemployment Insurance premiums, and Income Tax. It is proposed to use the monthly remittance as an auxiliary variable in the estimation of the SEPH variables. A previous study by Cotton(1987) explored the possibility of using the remittance data in this way and showed the prospect of substantial gains in efficiency.

The simulation study was conducted to identify the best sampling strategy and the best estimator among the ratio- and regression-type to meet the stated corporate concerns. Two subsidiary concerns guided the study:

- maintain the current reliability of the SEPH level estimates;
- minimize any changes to the current SEPH design and processing systems.

This report is structured as follows:

- Section 2: the current SEPH design;
- Section 3: objectives, constraints and key issues;
- Section 4: the methodology;
- Section 5: the results of the study;
- Section 6: conclusions and recommendations.

2. THE CURRENT SEPH DESIGN

The SEPH universe consists of approximately 700,000 employee reporting units (ERUs), from which the monthly sample of about 70,000 units is drawn. SEPH has a stratified, simple random sampling design with fixed sampling rates. There are 4 levels of stratification: (i) geography - Province/Territory of Canada; (ii) Standard Industrial Classification (SIC); (iii) size-group, which is determined by the total number of paid employees; and, (iv) the Take-All/Take-Some classification, which indicates a combination of both size and complexity of structure. There are four size-groups defined as follows:

Size-group 1: 0 - 19 employees;
Size-group 2: 20 - 49 employees;
Size-group 3: 50 - 199 employees;
Size-group 4: 200 or more employees.

A Take-All stratum is sampled with certainty and includes all units size-group 4 and all multi-ERU companies regardless of size-group. (This is the old definition. The new definition does not include small multi-ERUs whose aggregated sizes are less than 200 employees. The data used in the study are old and thus, the old definition is given.) Take-Some units are sampled with probability less than one.

The basic building block of SEPH is the cell - a stratum defined by 3-digit SIC (SIC3), province/territory (PROV), and size-group. There are 13,488 such cells, about half of which are empty and many of which are very small. These cells are further stratified by the Take-All/Take-Some classification.

The Take-Some ERUs in each cell are sampled with fixed sampling rates which are determined by the level of precision sought (a coefficient of variation of 3% at industry division x PROV level of aggregation, where the industry division (IND) is defined as a collection of closely related SIC3s; there are 16 INDs). The Take-Some portion of the sample is rotated each month by replacing 1/12 of the ERUs. The units rotated out of the sample are not eligible for re-selection for at least 12 months.

SEPH uses the expansion estimator. The sample allocation scheme and other details of the SEPH methodology are given in Schiopu-Kratina and Srinath (1986).

3. OBJECTIVES, CONSTRAINTS AND KEY ISSUES

In the light of the subsidiary concerns noted in the introduction and owing to some difficulties of linking the remittance data to multi-ERU companies, Take-All units were excluded from the target population for the study. Size-group 3 was also excluded because only a small sample size reduction was expected owing to the high sampling rate in this group.

It was obvious from a preliminary investigation that a significant sample size reduction would not be possible if the new estimator were applied at the cell level. Hence, it was decided to collapse SIC3s into 2-digit SICs (SIC2) in order to elevate the application level. Assuming that the regression estimator was to be applied at the SIC2 x PROV level, a rough estimate of a 10,000 sample size reduction from the Take-Some portion of size-groups 1 and 2 was obtained.

Several important issues arise. These need to be resolved if the new methodology is to be successfully implemented and the objective realized. These key issues are listed as follows:

1) The Quality of the Remittance Data
The remittance data have data quality problems owing to various reasons, such as, one time bulk payment, late payment and processing error, etc. It was found that the use of two- or three-months' average remittance, instead of the monthly remittance, increases the correlation with a SEPH variable (Cotton, 1987). However, the average is not robust against outlying values. This indicates the need to use a robust technique.

2) Selection of an Appropriate Estimator
An estimator which is efficient enough to effect the 10,000 sample size reduction is needed. We confined the choice among ratio- and regression- type estimators. Since there is the size stratification within the level where an estimate is required, it should be determined whether combining of the size strata in computing the estimate is beneficial or not.

3) The Level of Application
If the sample size reduction of 10,000 cannot be achieved by collapsing at the SIC2 level, then perhaps collapsing at a higher level, e.g., IND level may be necessary especially in the small provinces and territories.

4) Small Area Estimation
The level of application of a proposed estimator should be above the cell (i.e., SIC3 x PROV), but estimates are still required at the cell level. For this purpose, we have to use an appropriate small area estimation technique.

4. METHODOLOGY OF THE STUDY

The basic methodology used for the study is a Monte Carlo simulation. In order to conduct a simulation study, we needed population data sets which resembled the SEPH universe, so that the results are applicable to SEPH. Four population data sets, each containing data for 2 consecutive months, were created for this purpose by using the SEPH sample data. Two months of data were necessary in order to estimate the month-to-month change, as well as the monthly level. The 4 populations are classified according to the average size of combined strata of size-groups at the level of application. They are described below:

1) The Small-Size Population 1 (SSP1)
It was generated using the SEPH data for the wholesale trade industry (IND = 61) from all provinces excluding the two territories.

2) The Small-Size Population 2 (SSP2)
Again SEPH sample data were used directly to generate this population. The data from the two territories for all industries were used. The level of application in this population was the industry division rather than the 2-digit SIC as in the other three populations.

3) The Medium-Size Population (MSP)
The SEPH data from the 10 provinces for the building construction industry (IND = 41) were used as this population.

4) The Large-Size Population (LSP)
There was no SEPH sample data which could be used directly as a large-size population. Hence, it was generated artificially using a multivariate technique with population parameters obtained from the SEPH data from Quebec, Ontario and Western Provinces for the commercial service industry (IND = 87). It has no size-groups. See Lee(1989) for details.

Two hundred replicates of simple random samples were drawn from each size-group at the application level (i.e., SIC2 or IND x PROV) from SSP1, SSP2 and MSP and 100 replicates from LSP.

These populations provide a variety of combinations of population and sample sizes, correlation, industry and geography. The following table gives the characteristics of the populations.

Table 1: Description of the Population Data Sets Used in the Simulation Study

	SSP1	SSP2	MSP	LSP
Data Source				
Industry	61	All	41	87
Province[1]	10-59	60,61	10-59	24-59
No. of Cases[2]	60	50	40	72
Ave. Pop. Size				
Size Grp 1 & 2	56	46	230	5004
Ave. Sample Size	16	9	46	65
Ave. Correlation[3]				
EMP[4]	0.72	0.69	0.85	0.63
GRP	0.83	0.77	0.89	0.78
HRS	0.74	0.69	0.84	0.66
Generation Method	Sample Data	Sample Data	Sample Data	Artificial
Application Level	SIC2	IND	SIC2	SIC2
No. of Replicates	200	200	200	100

Note 1: The province code definitions are: 10 = Newfoundland; 11 = Prince Edward Island; 12 = Nova Scotia; 13 = New Brunswick; 24 = Quebec; 35 = Ontario; 46 = Manitoba; 47 = Saskatchewan; 48 = Alberta; 59 = British Columbia; 60 = Yukon; 61 = North Western Territory.

Note 2: The case is defined by 2-digit SIC x PROV for SSP1, MSP and LSP, and IND x PROV for SSP2 which is the level of application.

Note 3: The correlation is between a SEPH variable and remittance data.

Note 4: EMP, GRP and HRS are abbreviations for Employment, Gross Payroll and Hours, respectively.

4.1 Estimators Considered

The estimators used in the study are: (i) the separate and combined ratio estimators; (ii) Mickey's unbiased ratio estimator; (iii) the separate and combined regression estimators (for the definitions of these estimators, see Cochran, 1977) - the combining is over size-groups 1 and 2. The expansion estimator was used for comparison, e.g., for obtaining the relative efficiency of the other estimators.

4.2 Small Area Estimators

Nine small area estimators were studied for estimation at the SIC3 x PROV level: 2 synthetic estimators, 2 composite estimators, four empirical best linear unbiased predictors (EBLUP), and the expansion estimator for comparison. Short descriptions of these estimators are given in the following table and the models from which these estimators are derived are listed in the Appendix. For detailed definitions, see Choudhry and Rao (1988).

No.	Name	Description
1	Expansion	Blow-up estimate or survey estimate
2	Synthetic 1	Based on Model 1 with constant β
3	Synthetic 2	Based on Model 2 with constant β
4	Composite 1	Linear combination of Nos. 1 and 2
5	Composite 2	Linear combination of Nos. 1 and 3
6-9	EBLUP 1-4	Based on Models 1-4

In general, these small area estimates for SIC3s in a given SIC2 will not be additive to the SIC2 level estimate. In order to obtain the additivity, a benchmarking procedure was applied as given in the Appendix.

4.3 Improving the Remittance Data Quality

Simple robust techniques such as medians and trimmed means of 3, 5 ,7, or 9 months' remittances were studied. These methods usually led to further increased correlations with the SEPH variables compared to the 2- or 3-months' average remittances. All these robust methods produced similar results. For simplicity and ease of computation, the 3-months' median remittance was chosen. In the following table, the average correlation of EMP with the 3-months' median remittance is compared with those for raw (untreated monthly) and the 3-months' average remittances for the two months, October and November, 1987:

Method	Oct	Nov
Raw	0.75	0.68
3-Months' Average	0.77	0.76
3-Months' Median	0.79	0.78

It is interesting to observe that the correlation with raw remittance in November is much lower than that in October but the correlations with the 3-months' average and median are at the same level in both months. The margin of improvement seems to be greater when there are more serious data quality problems in the remittance. Moreover, the methods tend to stabilize the correlation over time.

5. THE RESULTS OF THE STUDY

The various competing estimators were compared with the usual expansion estimator. The relative efficiency (REFF) is used as criterion for rating the estimators. It is defined as:

$$\text{REFF} = 100\frac{\text{MSE of Expansion Estimator}}{\text{MSE of Alternative Estimator}}.$$

The number of cases for which an estimator achieved a gain in efficiency was also used to rate the estimator. Here, the 'case' means a stratum defined by SIC2 x PROV where sampling takes place and the estimators are applied.

5.1 The Results for Level Estimates

Tables 2a and 2b show the average REFF of the alternative estimators for Employment with untreated remittance data. As expected, their performances are dependent on the sizes of population and sample, and the correlation. In SSP1, all alternative estimators showed loss in about 50% of cases and their average REFFs range from 66 to 133. In LSP, ratio and Mickey's estimators are only slightly better than in SSP1 even though sample sizes are much bigger (average sample size is 65 vs. 16 in SSP1). However, the regression estimator performed very well in LSP with the average REFF of 202 and 82% of cases showing gains. The rather poor performances of ratio and Mickey's estimators indicates that the intercept term of the linear regression line is significantly different from zero. No attempt was made to confirm this by statistical testing since this seems to be obvious. In MSP, all alternative estimators did quite well for obvious reasons, namely large correlations and sample size.

Table 2a: Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for EMP with Untreated Remittance

Population	RATIO		REGRESSION		MICKEY's
	SEP.	COM.	SEP.	COM.	
SSP1:					
% of Gains	30	43	52	60	40
Ave. REFF	74	113	115	133	66
MSP:					
% of Gains	75	78	98	98	68
Ave. REFF	141	176	221	240	132

Table 2b: Average Relative Efficiency of the Alternative Estimators w.r.t. the Expansion Estimator for EMP with Untreated Remittance

Population	RATIO	REGRESSION	MICKEY's
SSP2[1]:			
% of Gains	42	38	34
Ave. REFF	246	159	232
LSP[2]:			
% of Gains	56	82	40
Ave. REFF	147	202	132

Note 1: Pooled estimators were used in SSP2.
Note 2: No size-group was defined in LSP.

Among the competitors considered here, the regression estimator is clearly the best for the populations except for SSP2. The combined ratio and regression estimators have greater efficiencies than their separate counterparts. Based on the criteria for rating the estimators, the combined regression estimator is the best for the three populations; SSP1, MSP and LSP.

The results for SSP2 are quite different from those for the other populations. The ratio estimator was the most efficient with average REFF of 246 compared with 159 for the regression estimator. This may be mainly due to the fact that the level of application of the estimators is the industry division and not SIC2 as for other populations. It would seem that collapsing of SIC2s into industry division makes the population data follow more closely the super population model $y_i = \beta x_i + e_i$, $E(e_i) = 0$ and $V(e_i) = x_i\sigma^2$, for which the ratio estimator is the best.

It is somewhat surprising to observe that the Mickey's estimator performed considerably worse than the ratio estimator. This means that the unbiasedness of the Mickey's estimator has a high price tag. This observation, however, pertains only to the type of populations used in this study. Hidiroglou and Choudhry (1989) obtained different results with other populations which showed very close performances of the two estimators in terms of MSE.

The results for Gross Payroll are much more favorable to the alternative estimators (for example, the average REFF of the regression estimator for Gross Payroll in LSP is 488 vs. 202 for Employment and the percentage of cases showing gains is 92 vs. 82) owing to much larger correlations. For Hours, the results are very similar to those for Employment.

When treated remittance data were used, the performance of the combined regression estimator was much better than with untreated data as shown in Table 3. The improvements are more prominent in SSP1 than MSP, which implies that the use of three-months' median enhances the correlations more in SSP1 than MSP in which the correlations are already quite high (see Table 1).

Table 3: Comparison of the Average REFF of the Combined Regression Estimator Using the Raw and the Treated Remittance

Population		EMP		GRP	
		Raw	Trtd	Raw	Trtd
SSP1:	% of Gains	60	77	82	92
	Ave. REFF	133	177	313	527
MSP:	% of Gains	98	98	100	98
	Ave. REFF	240	260	391	474

5.2 The Change Estimate

The change estimate is also important for SEPH. It is usually obtained by subtracting the previous month's level estimate from the current month's level estimate. The change estimate obtained by using the combined regression estimator with treated remittance data is compared with that by the expansion estimator in Table 4. The average REFFs are considerably smaller than those for the level estimates. This means that the combined regression estimator for change is not as efficient as for level and thus, a combined regression estimator which maintains the current reliability for level with reduced sample size will not keep the same reliability for change.

Table 4: Average REFF of the Change Estimate using the Combined Regression Estimator with Treated Remittance

Population		EMP	GRP
SSP1:	% of Gains	63	60
	Ave. REFF	181	105
MSP:	% of Gains	95	90
	Ave. REFF	218	172

5.3 Bias and Aggregated Estimates

The ratio and regression estimators are biased. The bias problem could be serious when it is accumulative at a higher level aggregation. About 70% of cases show a positive bias. The direction of bias persists when SIC2 estimates are aggregated to give national IND estimates. Consequently, the average REFFs of the aggregated estimates are substantially decreased as shown in Table 5. The nonzero bias for the expansion estimator in the table is due to the Monte Carlo error. The bias for the aggregated estimates of the ratio estimator is larger than that for the regression estimator.

Table 5: Aggregated[1] Estimates for Medium Size Population with Treated Remittance

Summary Statistic	Variable	Expansion	Comb. Ratio	Comb. Regression
Rel. Bias (%)	EMP	0.6	1.4	1.0
	GRP	0.7	0.8	0.8
	HRS	0.7	1.4	1.1
Rel. RMSE	EMP	3.6	2.9	2.6
	GRP	4.5	2.4	2.2
	HRS	3.9	3.2	2.7
Ave. REFF	EMP	100	125	200
	GRP	100	315	377
	HRS	100	156	215

Note 1: Aggregated over two 2-digit SICs and ten Provinces.

5.4 Small Area Estimation for 3-Digit SICs

Table 6 shows the results for the 9 small area estimators for SIC3 level estimates, benchmarked and unbenchmarked. Three criteria are used to compare their performances: (i) the Absolute Relative Bias (ARB) defined as 100(Absolute Bias / Population Total); (ii) the Absolute Relative Error (ARE) defined as 100(Absolute Error / Population Total); and (iii) REFF with respect to the expansion estimator. We will discuss the unbenchmarked results first.

Table 6: The Performances of the Small Area Estimators for EMP (averaged over 179 small areas)

Estimator	Unbenchmarked			Benchmarked		
	ARB	ARE	REFF	ARB	ARE	REFF
Expansion	5	68	100	6	70	100
Synthetic 1	26	28	248	26	28	260
Synthetic 2	***1	***	0	26	28	263
Composite 1	19	34	188	18	38	168
Composite 2	***	***	2	133	176	26
EBLUP 1	171	180	6	17	27	201
EBLUP 2	***	***	1	41	66	73
EBLUP 3	17	25	248	17	26	229
EBLUP 4	***	***	2	25	41	67

Note 1: This indicates that the number exceeds 1000.

Three estimators stand out among the unbenchmarked estimators. They are Synthetic 1, composite 1 and EBLUP 3. EBLUP 3 is the best and is followed closely by Synthetic 1. The estimators based on the models with the error variance proportional to x are generally good except EBLUP 1. All others did very poorly and surprisingly much worse than the expansion estimator. All of them are based on the models with the error variance proportional to x^2. The assumption on the error variance of the model seems to be the most important factor for selecting an appropriate model. The model for EBLUP 1 has a variable β and no intercept term while the model for EBLUP 3 has a constant β and nonzero variable intercept. These results seem to indicate that the model with intercept term with the error variance proportional to x is more appropriate. The synthetic estimator based on this model could be better than Synthetic 1.

The benchmarked results show a very different picture. The estimators which performed poorly without benchmarking were improved tremendously. The most dramatic improvement of performance was shown by Synthetic 2 which becomes equal to Synthetic 1. In fact, they are exactly the same under benchmarking (see the definitions of the benchmarked Synthetic 1 and 2 in Appendix). A small discrepancy in these estimators' REFFs in Table 6 is due to rounding error. Four estimators stand out in this case. These are Synthetic 1 and 2 and EBLUP 1 and 3. EBLUP 1 and 3 are superior to Synthetic 1 and 2 in terms of ARB and ARE but vice versa in terms of REFF. EBLUP 1 and 3 are less biased but more variable than Synthetic 1 and 2.

Without benchmarking, EBLUP 3 is the best. With benchmarking, EBLUP 3 can still be a good choice but Synthetic 1 (or 2) could be a better choice for its simplicity.

6. Conclusions and Recommendations

Conclusions are summarized as follows:

(1) In general, the ratio, regression and Mickey's estimators are superior to the expansion estimator. But it is risky to use these estimators when sample sizes are small unless the correlation is very high.

(2) The regression estimator is generally better than the ratio which is better than the Mickey's. The difference, however, gets smaller as population and the sample sizes get smaller.

(3) Collapsing at industry division level favors the ratio estimator.

(4) The combined estimators are generally better than the separate ones.

(5) Biases of the ratio and regression estimators are generally positive and thus accumulative at higher levels of aggregation. The bias of the ratio estimator is larger than that of the regression.

(6) The average REFFs of the ratio and regression estimators at higher level aggregation are substantially eroded by accumulated bias.

(7) Using the median of three-months' remittances improves and stabilizes correlation over time.

(8) The performances of the ratio and regression estimators for estimating change are not as good as for estimating level.

(9) Selecting a proper model is very important for small domain estimation. The model with error variance proportional to x seems to be appropriate.

(10) The simple synthetic estimator seems to be the best among those studied for small area estimation.

Based on these conclusions, we recommend the combined regression estimator at the SIC2 x PROV level. However, application of the estimator should be selective, in that we should carefully examine the population and sample sizes and correlation simultaneously to see whether the new estimator can bring about a gain and apply the new estimator only for strata where some efficiency gains are feasible.

The amount of sample size reduction should depend on the REFF of the combined regression estimator at the level of application. The larger the relative efficiency, the larger the sample size reduction. Applying regression analysis to the results of the study, we obtained a formula which gives the REFF as a function of the population and sample sizes and the correlation coefficient. Based on this formula, a sample size reduction scheme was established and from this we obtained an estimate of sample size reduction which was very close to the goal of 10,000. This was obtained without collapsing of SIC3s into a higher level than SIC2. Hence, we recommend that the level of application of the new estimator should be SIC2 x PROV throughout the whole country.

For small area estimation at SIC3 x PROV level, we recommend the use of Synthetic 1.

APPENDIX

The models used to derive the small area estimators studied in this paper are given in the following:

Model 1:

$$y_{ij} = \beta_i x_{ij} + e_{ij}, \quad \beta_i = \beta + \alpha_i, \quad V(e_{ij}) = x_{ij}\sigma^2.$$

Model 2:

$$y_{ij} = \beta_i x_{ij} + e_{ij}, \quad \beta_i = \beta + \alpha_i, \quad V(e_{ij}) = x_{ij}^2\sigma^2.$$

Model 3:

$$y_{ij} = \beta x_{ij} + \alpha_i + e_{ij}, \quad V(e_{ij}) = x_{ij}\sigma^2.$$

Model 4:

$$y_{ij} = \beta_i x_{ij} + \alpha_i + e_{ij}, \quad V(e_{ij}) = x_{ij}^2\sigma^2.$$

For all these models, it is assumed that $E(\alpha_i) = 0$ and $E(e_{ij}) = 0$. The subscript i stands for the i-th small area and the subscript j for the j-th unit in the i-th small area.

The benchmarked small area estimators by the combined regression estimator are defined as follows:

Let $\hat{Y}$ be the combined regression estimate at a given SIC2 X PROV and let $\hat{Y}_{s_1}, \ldots, \hat{Y}_{s_k}$ be the small area estimates for the SIC3s in the SIC2. Then the benchmarked small area estimate $\tilde{Y}_{s_i}$ is defined as:

$$\tilde{Y}_{s_i} = \frac{\hat{Y}_{s_i}}{\hat{Y}_{s.}}\hat{Y}, \quad \hat{Y}_{s.} = \sum_{j=1}^{k} \hat{Y}_{s_j}.$$

For Synthetic 1 and 2, it reduces to:

$$\tilde{Y}_{s_i} = \frac{X_{s_i}}{X_{s.}}\hat{Y}$$

where $X_{s_1}, \ldots, X_{s_k}$ are the total remittances of the SIC3s and $X_{s.}$ is their sum.

ACKNOWLEDGEMENT

The authors would like to thank Mike Hidiroglou and Dave Dolson of Statistics Canada for their helpful comments.

REFERENCES

Choudhry, H., and Rao, J.N.K., (1988). Evaluation of Small Area Estimators: An Empirical Study. Presented at the International Symposium on Small Area Statistics, New Orleans.

Cochran, W.G., (1977). Sampling Techniques; 3rd Edition. New York: John Wiley & Sons.

Cotton, C., (1987). PD 7 Remittances and SEPH. Technical Report, Business Survey Methods Division, Statistics Canada.

Hidiroglou, M.A., and Choudhry, H., (1989). Sampling and Estimation Methodology for Sub-Annual Surveys at Statistics Canada. Technical Report, Business Survey Methods Division, Statistics Canada (submitted to Survey Methodology).

Lee, H., (1989). Generation of an Artificial Population Which Resembles Multi-Variate Business Survey Data. Technical Report, Business Survey Methods Division, Statistics Canada (under preparation).

Mickey, M.R. (1959). Some Finite Population Unbiased Ratio and Regression Estimators. Journal of American Statistical Association, Vol. 54, pp. 594-612.

Schiopu-Kratina, I., and Srinath, K.P., (1986). The Methodology of the Survey of Employment, Payroll and Hours. Technical Report, Business Survey Methods Division, Statistics Canada.